Proc. Assoc. Advmt. Anim. Breed. Genet. 22:425-428

AGREEMENT AMONG GWAS RESULTS FROM DIFFERENT STATISTICAL METHODS AS A STRATEGY TO INCREASE THE POWER OF QTL DETECTION

T.P. Melo[1], B.F. Garcia Neto[1], M.R.S. Fortes[2], L.G. Albuquerque[1,4] and R. Carvalheiro[1,4]

[1] School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal, SP, Brazil
[2] School of Chemistry and Molecular Biosciences, University of Queensland, Queensland, Australia
[3] Queensland Alliance for Agriculture and Food Innovation (QAAFI) Institute, University of Queensland, Queensland, Australia
[4] National Council for Scientific and Technological Development (CNPq), Brasília, DF, Brazil

SUMMARY

The power of true positive associations in GWAS for traits affected by many QTL is generally low. This and other unfavorable scenarios pose a problem for the detection of true QTLs, which may lead to false positive associations. The aim of this study was to evaluate if combining the results of different statistical methods may increase the power to detect QTL. We simulated a polygenic trait, with known QTL positions. GWAS was performed using the WssGBLUP and BayesC methods, in a total of 8 different analyses, varying the assumptions of the SNP effects and the phenotypic data used. The results showed that as the number of analyses that a window was detected as important increased, so did the probability of that window containing a true QTL. Windows identified in 7 or 8 analyses were able to detect just some (60.5%) of the true QTL. Windows detected in at least 5 analyses captured 96% of the true QTL, but included some false positives (10.8%). Further studies are recommended, simulating traits with different genetic architectures, under different population structures, to evaluate the reproducibility of the present results.

INTRODUCTION

QTL detection remains a challenge in animal breeding, especially for lowly heritable complex polygenic traits. Under this scenario, Genome Wide Association Studies (GWAS) may present low power or high number of false positives, depending on the significance threshold adopted. Many statistical methods to perform GWAS are available (Meuwissen *et al.* 2001; Habier *et al.* 2011; Wang *et al.* 2012, and others), however their efficiency will depend on several factors such as the genetic architecture of the trait and the modeling assumptions related to the markers effects. Furthermore, other factors such as the linkage disequilibrium and the amount of phenotypic and genotypic information available may also affect the ability of QTL detection (Melo *et al.* 2016).

When a genome region is detected as important by many statistical methods, the evidence that this region harbours a true QTL is supposedly increased (Legarra *et al.* 2015). The aim of this study was to evaluate if the number of statistical methods for which a region is considered to be significant is associated with the power of QTL detection, for a simulated lowly heritable complex trait.

MATERIAL AND METHODS

Simulation. QMSim software (Sargolzae & Schenkel 2013) was used to simulate a trait with heritability and phenotypic variance equal to 0.14 and 1, respectively. A historical population, with constant size of 1,000 animals (500 males: 500 females), was simulated for 1,000 generations. The population size was then decreased until it reached 200 animals (100 females), over another 2,020 historical generations, producing a bottleneck effect and, as a consequence, genetic drift and linkage disequilibrium. The 200 animals from the last generation of the historical population were selected

as the founders of an expansion population, simulated over 6 generations. In this expansion process, the number of females grew exponentially and each dam had five offspring in each generation, totaling 16,000 animals (8,000 females) at the end of the expansion process. A total of 240 males and 6,000 females from the last expansion population were randomly selected to be the founders of the selection population. The selection was performed over another 15 generations, using a replacement rate of 20% for males and females, based on estimated breeding values. Phenotypic information of the females (≈45,000) from all generations of the selection population and 2,000 randomly selected genotypes from females of the last three generations were used to perform the GWAS. This small proportion of genotyped animals was chosen to mimic a common situation. The simulated genome had a length of 2,333 cM, 735,293 markers and 7,000 QTLs. The average number of markers and QTLs per chromosome was 16,782 and 158, respectively, randomly distributed over 29 autosomes. It was assumed that QTLs explain 100 % of genetic variance. QTL allele effects were sampled from a gamma distribution with a shape parameter of 0.4, and the phenotypes were generated summing the effects of 1,000 randomly selected segregating QTLs to an error term sampled from a normal distribution with zero mean and variance of 0.86. Ten replicates of the simulation process were performed. More details about the simulation are available in Melo *et al.* (2016).

Statistical methods. Two statistical methods were used to perform the GWAS, namely weighted single-step GBLUP (WssGBLUP; Wang *et al.* 2012) and BayesC (Habier *et al.* 2011). The model adopted for WssGBLUP was: $\boldsymbol{y}=\boldsymbol{1\mu}+\boldsymbol{Z_a a}+\boldsymbol{e}$, where $\boldsymbol{y}$ is the vector of phenotypes, $\boldsymbol{\mu}$ is the overall mean, $\boldsymbol{a}$ is the vector of additive genetic effects, $\boldsymbol{1}$ is a vector of ones, $\boldsymbol{Z_a}$ is an incidence matrix relating the phenotypes to $\boldsymbol{a}$, and $\boldsymbol{e}$ is the vector of residuals. The covariance between $\boldsymbol{a}$ and $\boldsymbol{e}$ was assumed to be zero and their variances were considered to be $H\sigma_a^2$ and $I\sigma_e^2$, respectively, where σ_a^2 and σ_e^2 are the direct additive and residual variance, respectively, $\boldsymbol{H}$ is the matrix which combines pedigree and genomic information (Aguilar *et al.* 2010), and $\boldsymbol{I}$ is an identity matrix. The SNP effects ($\hat{\boldsymbol{u}}$) were calculated as in Stranden & Garrick (2009): $\hat{u}=DP'[PDP']^{-1}a_g$, where $\boldsymbol{D}$ is a diagonal matrix that contains the weights for the SNPs, $\boldsymbol{P}$ is a matrix relating genotypes of each locus (coded as 0, 1 or 2 according to the number of copies of allele B) and $\boldsymbol{a_g}$ is a vector with the estimated breeding values of genotyped animals. $\boldsymbol{D}$, $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{u}}$ were iteratively recomputed over three iterations. In the first iteration (w1), the diagonal elements of $\boldsymbol{D}$ (d_i) were assumed to be 1 (i.e., the same weight for all markers). For the subsequent iterations (w2 and w3), d_i was calculated as: $d_i=\hat{u}_i 2p_i(1-p_i)$, where $\hat{u}_i$ is the allele substitution effect of the i^{th} marker, estimated from the previous iteration, and p_i is the allele frequency of the second allele of the i^{th} marker. The WssGBLUP was adopted using two sets of data, one considering all available phenotypic information (SI; n=45,000) and another considering phenotypes just from genotyped animals (SII; n=2,000). The three different weights for the SNPs (w1 to w3) and the two sets of data (SI and SII) resulted in six different solutions for the SNP effects obtained under the WssGBLUP method. BayesC was applied under the model: $\boldsymbol{y}=\boldsymbol{1}\mu+\sum_{i=1}^{n}\boldsymbol{g_i b_i \delta_i}+\boldsymbol{e}$, where $\boldsymbol{y}$, $\boldsymbol{1}$, $\boldsymbol{\mu}$ and $\boldsymbol{e}$ are as previously described, $\boldsymbol{g_i}$ is the vector with the genotype of the animals for the i^{th} SNP, $\boldsymbol{b_i}$ is the vector containing the allele substitution effect of the i^{th} SNP and δ_i is an indicator variable (0, 1), with parameter π, where π is the fraction of SNPs not included in the model. Two π values were used, 0.99 or 0.999. The genotypes were coded as AA = 0, AB =1 and BB = 2. In summary, a total of 8 analyses were performed: WssGBLUP SI and SII (w1, w2 and w3), and BayesC (π=0.99 and π=0.999). The GWAS results were compared based on the proportion of variance explained by SNPs within consecutive 1Mb windows. For each analysis, the top 20 marker windows, which explained the greatest proportion of genetic variance, were identified and their locations were contrasted with the true QTL position. A true QTL was considered to be mapped when a top marker window was located no more than 1 Mb from a true QTL that explained at least 1% of the genetic additive variance.

RESULTS AND DISCUSSION

The simulation process resulted on average in 16.7 (±2.8) QTLs explaining 1% or more of the genetic variance. Together, the true QTLs explained on average 29.7% (±4.9) of the genetic variance, with the most important QTL explaining on average 5.1% (±2.4). The different analyses presented poor ability to map the QTLs. Individually, they were able to identify between 5.4% (WssGBLUP; SII; w3) and 17.4% (WssGBLUP; SI; w2) of the true QTLs. The power of QTL mapping increased when a window was detected as significant by different analyses (Figure 1). The percentage of true associations increased along with the number of analyses, reaching a maximum of 100% (i.e. 0% of false positives) when a window was identified as important by 7 or 8 analyses. Although presenting just true associations, windows identified in 7 or 8 analyses were able to detect just part (60.5%) of the true QTL, since some QTL were not mapped by 7 or 8 analyses, however this percentage is still high compared with the worse scenario (1.7%) in which a window was detected just by 1 analysis. The maximum percentage of true QTLs identified was observed when a window was considered as important in 5 analyses, where 96% of the true QTL were identified. This scenario (5 analyses) presented, however, 10.8% of false positive associations (Figure 1).

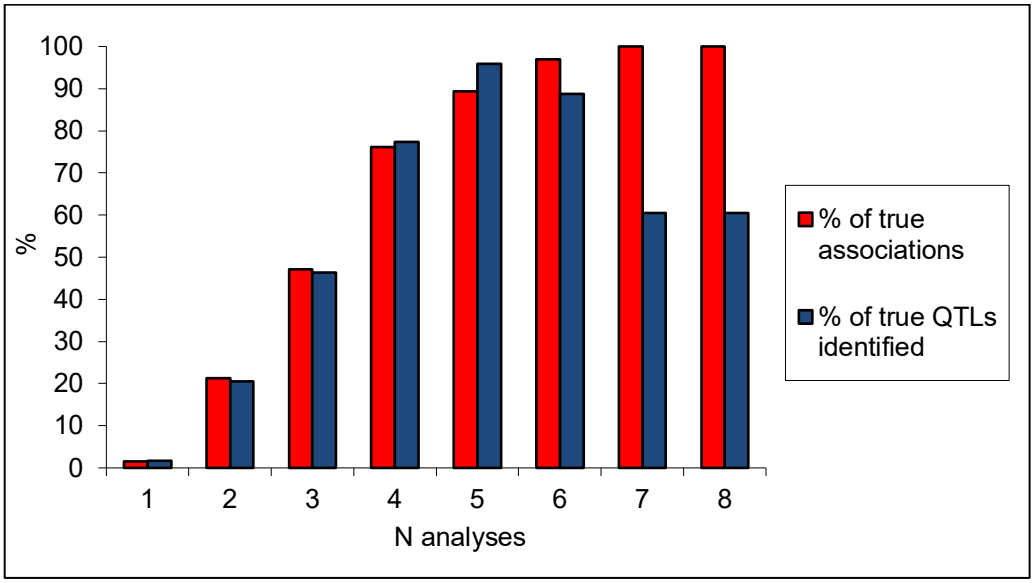

Figure 1. Percentage of true associations and of true QTL detected according to the number of analyses in which a marker window was identified as important

Our results are in accordance with Legarra *et al.* (2015), who recommended using different methods to map QTL more efficiently, arguing that no method is markedly more powerful, being dependent on the genetic architecture of the trait. Van den Berg *et al.* (2013), assessed through simulation the power of BayesC and BayesCπ to detect QTL, and also observed poor ability to detect QTL for lowly heritable complex traits. Unfortunately, the authors did not test if the agreement between results of the different methods/analyses increased the power of QTL detection.

Although our simulation study did not cover all factors affecting the QTL detection in real complex traits, the results provide evidence that the agreement among results from different statistical GWAS methods may be a feasible strategy to map QTL more precisely, especially for lowly heritable polygenic traits. Further studies may investigate the optimal number and

combination of statistical methods, under different scenarios of heritability, number of genotyped animals family structure, effective population size, genetic architecture and considering other definitions of true QTLs, which would result in improved power of QTL detection.

In conclusion, our simulation approach demonstrated that agreement among GWAS results from different statistical methods can be used as a strategy to increase the power of QTL detection. This is a promising approach in the context that genomic selection can benefit from identification of true QTL (Pérez-Enciso *et al*. 2015). Our future proposition is to apply these methods to field data collected on beef cattle farms, targeting complex traits.

REFERENCES

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. and Lawlor, T.J. (2010) *J. Dairy Sci.* **93**: 743.

Habier D., Fernando R.L., Kizilkaya K., Garrick D.J. (2011) *BMC Bioinformatics* **12**: 1.

Legarra A., Croiseau P., Sanchez M.P., Teyssèdre S., Sallé G., Allais S., Fritz S., Moreno C.R., Ricard A., Elsen J-M. (2015) *Genet Sel Evol*. **47**: 1.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) *Genetics* **157**: 1819.

Melo T.P., Takada L., Baldi F., Oliveira H.N., Dias M.M., Neves H.H.R., Schenkel F.S., Albuquerque L.G., Carvalheiro R. (2016) *BMC Genetics* **17**: 1.

Pérez-Enciso M., Rincón J.C., Legarra A. (2015) *Genet Sel Evol*. **47**: 43.

Sargolzaei, M. and Schenkel, F.S. (2013) QMSim: User's Guide. p.77.

Stranden I., Garrick D.J. (2009) *J Dairy Sci*. **92**: 2971.

Van den Berg I., Fritz S., Boichard D. (2013) *Genet Sel Evol*. **45**: 1.

Wang H., Misztal I., Aguilar I., Legarra A., Muir W.M. (2012) *Genet Res.* **94**: 73.